Processus de Décisions Markoviens (MDP) et application

François Delarue

Rémi Catellier

Table des matières

Chapitre 3

Modèle à N périodes

Dans tout ce chapitre, S et A sont supposés au plus finis pour simplifier, mais beaucoup des concepts exposés ici s'étendent aux cas dénombrables voire continus. Cela oblige à prêter attention aux problèmes d'intégrabilité (pour le cas dénombrable) et de mesurabilité (pour le cas continu).

1. Exposition du problème

On se donne maintenant une chaîne de Markov contrôlée $((X_n, \alpha_n)_{0\leq n\leq N}, (\alpha_n, \alpha_n)_{0\leq n\leq N-1})$ entre les instants 0 et N, et associée à une collection de matrices de transition contrôlées

$$\big(P(n,i,a,j)\big)_{n=0,\dots,N-1;i,j\in S;a\in A}.$$

1.1. Récompenses. On généralise le principe du chapitre 2 en considérant :

(1) $r(n,\cdot,\cdot) : S \times A \to \mathbb{R}$ fonction de récompense instantanée à l'instant n ;
(2) $g : S \to \mathbb{R}$ fonction de récompense terminale.

Définition 1.1. Etant donnée une chaîne de Markov contrôlée $(X_n, \alpha_n)_{0\leq n\leq N}$ entre les instants 0 et N, on appelle récompense totale moyenne la quantité :

$$\mathscr{R} := \mathbb{E}\bigg[\sum_{n=0}^{N-1} r(n, X_n, \alpha_n) + g(X_N)\bigg].$$

Les espaces S et A étant supposés finis, les espérances ci-dessus sont nécessairement bien définies.

Le but, pour le joueur, est de choisir $\alpha_0, \cdots, \alpha_{n-1}$ de façon à obtenir le meilleur $\mathscr{R}$ possible. Ceci n'est pas aisé en raison du caractère aléatoire de $\alpha_0, \cdots, \alpha_{n-1}$.

1.2. Lien avec le modèle à une période. Se pose la question de bénéficier des résultats du chapitre précédent, sur le modèle à 1 période.

L'idée majeure est d'appliquer les principes du Chapitre 2 sur la **dernière période**, c'est-à-dire entre les instants $N-1$ et N. Ceci nous conduit à décomposer la quantité $\mathscr{R}$ de la façon suivante :

$$\mathscr{R} = \mathbb{E}\bigg[\sum_{n=0}^{N-2} r(n, X_n, \alpha_n) + \Big(r(N-1, X_{N-1}, \alpha_{N-1}) + g(X_N)\Big)\bigg].$$

Pour rendre compte de l'intérêt de l'instant $N-1$, nous choisissons de conditionner par rapport à $\mathcal{F}_{N-1}$ dans l'écriture ci-dessus. Nous obtenons :

$$\mathscr{R} = \mathbb{E}\bigg[\sum_{n=0}^{N-2} r(n, X_n, \alpha_n) + \mathbb{E}\Big(r(N-1, X_{N-1}, \alpha_{N-1}) + g(X_N) \,|\, \mathcal{F}_{N-1}\Big)\bigg].$$

Nous pouvons interpréter cette identité de la façon suivante : à l'instant $N-1$, toutes les récompenses jusqu'à l'instant $N-2$ ont été observées ; sachant toutes les observations jusqu'à $N-2$, le problème est de fait de maximiser :

$$\mathbb{E}\Big(r(N-1, X_{N-1}, \alpha_{N-1}) + g(X_N) \,|\, \mathcal{F}_{N-1}\Big).$$

Ceci nous ramène, comme espéré, au modèle à une période.

1.3. Utilisation des matrices de transition. Nous réécrivons :

$$\begin{aligned}&\mathbb{E}\Big(r(N-1,X_{N-1},\alpha_{N-1})+g(X_N)\,|\,\mathcal{F}_{N-1}\Big)\\&=r(N-1,X_{N-1},\alpha_{N-1})+\sum_{j\in S}g(j)P(N-1,X_{N-1},\alpha_{N-1},j)\end{aligned}$$

Nous en déduisons le résultat suivant :

LEMME 1.2. *Donnons nous, pour tout $i\in S$, $a_{N-1}(i)$, maximiseur de la fonction*

$$a\in A\mapsto r(N-1,i,a)+\sum_{j\in S}g(j)P(N-1,i,a,j).$$

Alors,

$$\begin{aligned}&\mathbb{E}\Big(r(N-1,X_{N-1},\alpha_{N-1})+g(X_N)\,|\,\mathcal{F}_{N-1}\Big)\\&\leq r\big(N-1,X_{N-1},a_{N-1}(X_{N-1})\big)+\sum_{j\in S}g(j)P\big(N-1,X_{N-1},a_{N-1}(X_{N-1}),j\big).\end{aligned}$$

Puis,

PROPOSITION 1.3. *Donnons nous, pour tout $i\in S$, $a_{N-1}(i)$, maximiseur de la fonction*

$$a\in A\mapsto r(N-1,i,a)+\sum_{j\in S}g(j)P(N-1,i,a,j).$$

Alors,

$$\begin{aligned}\mathscr{R}\leq\mathbb{E}\Bigg[\sum_{n=0}^{N-2}r(n,X_n,\alpha_n)+\mathbb{E}\Big(r\big(N-1,X_{N-1},a_{N-1}(X_{N-1})\big)\\+g\Big(X_N^{X_{N-1},a_{N-1}(X_{N-1})}\Big)\,|\,\mathcal{F}_{N-1}\Big)\Bigg].\end{aligned}$$

Que dit ce resultat ? Pour espérer obtenir le meilleur coût moyen, ii est nécessaire de jouer à l'instant $N-1$ une stratégie maximisant la quantité en hypothèse. Par exemple, une façon de faire est de jouer $a_{N-1}(X_{N-1})$. L'état $X_N^{X_{N-1},a_{N-1}(X_{N-1})}$ désigne alors le nouvel état terminal étant donnée la nouvelle stratégie jouée à l'instant $N-1$. Celui-ci est obtenu en appelant la transition $P(N-1,X_{N-1},a_{N-1}(X_{N-1}),\cdot)$

1.4. Réduction à un problème à $N-1$ périodes. Dans le paragraphe précédent, nous avons appris comment jouer sur la dernière période. La question est maintenant d'itérer ce principe.

Pour cela, il est nécessaire de dire ce que devient g. Il n'est pas possible d'utiliser le même. La fonction g représente en effet une récompense à l'instant N. Il faut maintenant écrire une récompense à l'instant $N-1$.

On pose

$$U_N(i)=g(i),\quad i\in S,$$

puis

$$U_{N-1}(i)=\max_{a\in A}\left[r(N-1,i,a)+\sum_{j\in S}g(j)P(N-1,i,a,j)\right].$$

Réécrivons $\mathscr{R}$ avec U_{N-1} à l'aide du lemme suivant :

LEMME 1.4. *Le gain $\mathscr{R}$ est majoré par*

$$\mathscr{R}\leq\mathbb{E}\left[\sum_{n=0}^{N-2}r(n,X_n,\alpha_n)+U_{N-1}(X_{N-1})\right],$$

avec égalité si

$$\mathbb{P}\Big(\big\{\alpha_n \in A_{N-1}(X_{N-1})\big\}\Big) = 1,$$

où

$$A_{N-1}(i) = \text{argmax}_{a\in A}\Big[r(N-1,i,a) + \sum_{j\in S} g(j)P(N-1,i,a,j)\Big].$$

Démonstration. De la Proposition 1.3, nous rappelons

$$\mathscr{R} \leq \mathbb{E}\Big[\sum_{n=0}^{N-2} r(n,X_n,\alpha_n) + r(N-1,X_{N-1},\alpha_{N-1}) + \sum_{j\in S} g(j)P(N-1,X_{N-1},\alpha_{N-1},j)\Big].$$

Nous en déduisons

$$\mathscr{R} \leq \mathbb{E}\Big[\sum_{n=0}^{N-2} r(n,X_n,\alpha_n) + r(N-1,X_{N-1},\alpha_{N-1}) + U_{N-1}(X_{N-1})\Big],$$

qui est la première inégalité de l'énoncé.

Pour obtenir le deuxième résultat, nous supposons qu'il y a égalité dans l'identité ci-dessus. Alors nous obtenons que

$$\mathbb{E}\big[U_{N-1}(X_{N-1})\big] = \mathbb{E}\Big[r(N-1,X_{N-1},\alpha_{N-1}) + \sum_{j\in S} g(j)P(N-1,X_{N-1},\alpha_{N-1},j)\Big].$$

Ceci peut aussi se réécrire

$$\mathbb{E}\big[U_{N-1}(X_{N-1}) - r(N-1,X_{N-1},\alpha_{N-1}) + \sum_{j\in S} g(j)P(N-1,X_{N-1},\alpha_{N-1},j)\Big] = 0,$$

mais par construction, nous savons que la variable aléatoire à l'intérieur de l'espérance est nécessairement positive ou nulle. Comme l'espérance est nulle, il vient

$$\mathbb{P}\Big(\Big\{U_{N-1}(X_{N-1}) - r(N-1,X_{N-1},\alpha_{N-1}) + \sum_{j\in S} g(j)P(N-1,X_{N-1},\alpha_{N-1},j) = 0\Big\}\Big) = 1.$$

Mais, l'évènement à l'intérieur de la probabilité coincide avec celui de l'énoncé, à savoir

$$\begin{aligned}
&\Big\{U_{N-1}(X_{N-1}) - r(N-1,X_{N-1},\alpha_{N-1}) + \sum_{j\in S} g(j)P(N-1,X_{N-1},\alpha_{N-1},j) = 0\Big\}\\
&= \Big\{\max_{a\in A}\Big(r(N-1,X_{N-1},a) + \sum_{j\in S} g(j)P(N-1,X_{N-1},a,j)\Big)\\
&\quad = r(N-1,X_{N-1},\alpha_{N-1}) + \sum_{j\in S} g(j)P(N-1,X_{N-1},\alpha_{N-1},j)\Big\},
\end{aligned}$$

soit encore

$$\begin{aligned}
&\Big\{U_{N-1}(X_{N-1}) - r(N-1,X_{N-1},\alpha_{N-1}) + \sum_{j\in S} g(j)P(N-1,X_{N-1},\alpha_{N-1},j) = 0\Big\}\\
&= \big\{\alpha_n \in A_{N-1}(X_{N-1})\big\}.
\end{aligned}$$

Le résultat suit. □

2. Programmation dynamique et équation de Bellman

2.1. Fonction valeur. Le Lemme 1.4 suggère d'étudier un nouveau problème d'optimisation consistant à maximiser

$$\mathbb{E}\Big[\sum_{n=0}^{N-2} r(n,X_n,\alpha_n) + U_{N-1}(X_{N-1})\Big],$$

où la fonction U_{N-1} est donnée par

$$U_{N-1}(i) = \max_{a\in A}\left[r(N-1,i,a) + \sum_{j\in S} g(j)P(N-1,i,a,j)\right].$$

La définition suivante donne un principe itératif visant à étendre la construction de U_{N-1} à chacune des périodes antérieures :

Définition 2.1. On appelle fonction valeur associée à la maximisation du gain $\mathscr{R}$ la fonction

$$(n,i) \in \{0,\cdots,N\}\times S \mapsto U_n(i),$$

définie par récurrence (descendante) par :

$$U_n(i) = \max_{a\in A}\left[r(n,i,a) + \sum_{j\in S} U_{n+1}(j)P(n,i,a,j)\right],$$

pour $n \in \{0,\cdots,N-1\}$.

Dans cette construction itérative, on n'a rien fait d'autre que de propager le passage de U_N à U_{N-1} aux périodes antérieures. Cette construction itérative porte le nom d'**équation de Bellman**.

2.2. Programmation dynamique. La problème est maintenant de relier la construction donnée par la Définition 2.1 avec l'objectif d'origine, à savoir la maximisation de $\mathcal{R}$.

On obtient la généralisation du Lemme 1.4 sous la forme suivante :

Proposition 2.2. *Pour chaque instant $n \in \{0,\cdots,N-1\}$, le gain $\mathscr{R}$ est majoré par*

$$\mathscr{R} \leq \mathbb{E}\left[\sum_{k=0}^{n-1} r(k,X_k,\alpha_k) + U_n(X_n)\right],$$

la somme ci-dessus étant comprise comme nulle si $n=0$. Par ailleurs, l'inégalité ci-dessus est une égalité si

$$\mathbb{P}\left(\bigcap_{\ell\in\{n,\cdots,N-1\}} \{\alpha_n \in A_n(X_n)\}\right) = 1,$$

où

$$A_n(i) = \mathrm{argmax}_{a\in A}\Big[r(n,i,a) + \sum_{j\in S} U_{n+1}(j)P(n,i,a,j)\Big].$$

Démonstration. On effectue la preuve par récurrence descendante sur la valeur du paramètre n.

On sait déjà que le résultat est vrai lorsque $n = N-1$. Il s'agit du Lemme 1.4 (que nous cherchons à généraliser ici).

Supposons maintenant le résultat vrai pour un certain $n \in \{1,\cdots,N-1\}$. Nous obtenons que

$$\mathscr{R} \leq \mathbb{E}\left[\sum_{k=0}^{n-1} r(k,X_k,\alpha_k) + U_n(X_n)\right].$$

Nous pouvons donc introduire un nouveau critère de maximisation posée sur le modèle à n périodes. Nous posons

$$\mathscr{R}_n := \mathbb{E}\left[\sum_{k=0}^{n-1} r(k,X_k,\alpha_k) + U_n(X_n)\right].$$

(Avec cette notation, $\mathscr{R}$ n'est rien d'autre que $\mathscr{R}_n$.)

Nous appliquons maintenant le Lemme 1.4 à $\mathscr{R}_n$. Il vient (avec la même convention que précédemment pour la somme)

$$\mathscr{R}_n \leq \mathbb{E}\left[\sum_{k=0}^{n-2} r(k, X_k, \alpha_k) + U_{n-1}(X_{n-1})\right] = \mathscr{R}_{n-1},$$

avec égalité si

$$\mathbb{P}\Big(\big\{\alpha_{n-1} \in A_{n-1}(X_{n-1})\big\}\Big) = 1.$$

Comme $\mathscr{R}$ est inférieur à $\mathscr{R}_n$, cela démontre la première partie du résultat.

Examinons maintenant le cas d'égalité. Pour avoir $\mathscr{R} = \mathscr{R}_{n-1}$, nous devons avoir à la fois $\mathscr{R} = \mathscr{R}_n$ et $\mathscr{R}_n = \mathscr{R}_{n-1}$. Par récurrence, la première égalité est vérifiée si

$$\mathbb{P}\left(\bigcap_{\ell\in\{n,\cdots,N-1\}} \big\{\alpha_n \in A_n(X_n)\big\}\right) = 1.$$

De même la deuxième égalité est vérifiée si

$$\mathbb{P}\Big(\big\{\alpha_{n-1} \in A_{n-1}(X_{n-1})\big\}\Big) = 1.$$

On en déduit la conclusion en réunissant les deux conditions. □

Remarque 2.3. *La preuve repose sur le principe suivant : pour que $\mathscr{R}$ soit égal à $\mathscr{R}_{n-1}$, il est nécessaire que $\mathscr{R}$ soit égal à $\mathscr{R}_n$ et que $\mathscr{R}_n$ soit égal à $\mathscr{R}_{n-1}$. Ce principe porte le nom de programmation dynamique : pour maximiser la compense entre $n-1$ et N (ce que l'on fait en identifiant $\mathscr{R}$ à $\mathscr{R}_{n-1}$), il faut être optimal entre n et N (ce que l'on fait en identifiant $\mathscr{R}$ à $\mathscr{R}_n$) et optimal entre $n-1$ et n (ce que l'on fait en identifiant $\mathscr{R}_{n-1}$ et $\mathscr{R}_n$).*

On en déduit donc le théorème suivant :

Théorème 2.4. *La récompense optimale $\mathscr{R}^*$ est égal à $\mathbb{E}[U_0(X_0)]$ et est obtenue en jouant à chaque étape $n \in \{0, \cdots, N-1\}$ une stratégie α_n vérifiant $\mathbb{P}(\{\alpha_n \in A_n(X_n)\}) = 1$.*

Remarque 2.5. *Ce théorème donne une autre inteprétation de U_0. On peut écrire $U_0(i)$ comme la meilleure récompense, en partant de i, c'est à dire*

$$U_0(i) = \max \mathbb{E}\left[\sum_{n=0}^{N-1} r(n, X_n, \alpha_n) + g(X_N)\right],$$

le maximum étant pris sur les actions aléatoires $(\alpha_n)_{0\leq n\leq N-1}$ et la condition initiale X_0 étant prise comme égale à i.

Ce qui est remarquable, c'est que $U_0(i)$ a été obtenue par une récurrence n'impliquant que des problèmes d'optimisation déterministes ! Nous avons donc réduit la complexité du problème à l'aide de la programmation dynamique.

Exercise 1. Sur la modèle de la représentation de $U_0(i)$, donner une représentation de $U_n(i)$ pour chaque $n \in \{0, \cdots, N-1\}$. On sommera les coûts à partir de l'instant n.

2.3. Stratégies optimales.

Proposition 2.6. *Appelons, pour tout $n \in \{0, \cdots, N-1\}$ et tout $i \in S$, $a_n(i)$ un élément de $A_n(i)$. Alors la stratégie $(\alpha_n)_{0\leq n\leq N-1}$ obtenue en jouant, à chaque instant n, $\alpha_n = a_n(X_n)$ est optimale. Cette stratégie est en forme markovienne.*

Exercise 2. Montrer, pour une stratégie optimale comme donnée dans l'énoncé, que la suite $(U_n(X_n))_{0\leq n\leq N}$ est une martingale relativement à la filtration $(\mathcal{F}_n)_{0\leq n\leq N}$ sous-jacente.

On écrira l'identité $U_n(i) = r_n(i, a(i)) + \sum_{j\in S} U_{n+1}(j)P(n, i, a, j)$ sous la forme d'une espérance conditionnelle sachant $\mathcal{F}_n$ calculée sur l'évènement $\{X_n = i\}$.

Exercise 3. En choisissant $S = \{-1, 1\}$, $A = \{-1, 1\}$ $N = 1$, $r_0(s, a) = \frac{1}{2}a^2$, $g(s) = \frac{1}{2}s^2$, et $P(i, j, a) = \mathbf{1}_{\{j=a\}}$, étudier les stratégies optimales.

3. Exemples

EXERCISE 4. On considère $S = \{-1, 1\}$, $A = \{0, 1\}$, $r(s, a) = s^2 + a$ et $P(s, 1, 1 - s) = p$ et $P(s, 0, s) = 1$. On pose $g(s) = s$.

Ecrire la récurrence de Bellman.

DÉMONSTRATION. On pose $U_2(i) = i$, puis

$$U_n(i) = \max\big[1 + a + a\big(pU_{n+1}(1 - i) + (1 - p)U_{n+1}(i)\big) + (1 - a)U_{n+1}(i)\big].$$

□